Extracting Training Data from Large Language Models

Year: 2020 | *Citations: 2398* | *Authors: Nicholas Carlini, Florian Tramèr, Eric Wallace*

Abstract

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a training data extraction attack to recover individual training examples by querying the language model.

We demonstrate our attack on GPT-2, a language model trained on scrapes of the public Internet, and are able to extract hundreds of verbatim text sequences from the model's training data. These extracted examples include (public) personally identifiable information (names, phone numbers, and email addresses), IRC conversations, code, and 128-bit UUIDs. Our attack is possible even though each of the above sequences are included in just one document in the training data.

We comprehensively evaluate our extraction attack to understand the factors that contribute to its success. For example, we find that larger models are more vulnerable than smaller models. We conclude by drawing lessons and discussing possible safeguards for training large language models.